論文の内容の要旨

Bayesian Statistical Methods for Extending Bilingual Lexicon Using Comparable Corpora
(ベイズ統計による Comparable Corpora からの対訳ペアの獲得)

氏名　アンドラーデ　シルファ　ダニエル　ゲオルグ

Bilingual dictionaries can be automatically extended by new translations using comparable corpora.The general idea is based on the assumption that similar words have similar contexts across languages. This thesis suggest a new method which is distinguished from previous work, by mainly two aspects: First, it captures the relevant context using a novel Bayesian estimation of the Point-wise Mutual Information. Second, the context is defined not only by a bag-of-words, but additionally enriched by dependency tree information, which is mapped across unrelated languages. We provide an in depth analysis of the performance of our method and compare it to several previous baseline methods. Furthermore this thesis also shows how the importance of different dependency tree information can be learned in a Bayesian framework.

Comparable corpora are two corpora written in different languages,covering similar topics. Comparable corpora are not necessarily parallel, and therefore, can be easily created for various domains.On the other hand, although general dictionaries are often abundantly available, domain-specific dictionaries are rare, and expensive to be manually created.We use comparable corpora to find a translation of a certain word (query word), in the following way: In the first step, from the context of the query word, we

extract salient pivot words. Pivot words are words for which a translation is already available in the bilingual dictionary. In the second step, we match these pivots across languages to identify translation candidates for the query word.For extracting relevant pivot words we use a Bayesian estimation of the Point-wise Mutual Information. We then calculate a similarity score between the query word and a translation candidate, by using the probability that the same pivots are extracted for both the query word and the translation candidate. We extract pivot words in several context positions, namely, bag-of-words of one sentence, and the successors, predecessor and siblings with respect to the dependency parse tree. In order to make these context positions comparable across the unrelated languages Japanese and English, we use several heuristics to adjust the dependency trees appropriately. We demonstrate that our proposed method can significantly increase the accuracy of word translations when compared to previous baseline methods.

In the final part of our thesis we introduce a supervised method which appropriately weights each context position. This method is based on a generalization of the cosine similarity: it performs a linear transformation of the context vectors using a specified matrix, before calculating the cosine similarity between them. The optimal matrix is expressed in a Bayesian probabilistic model and learned using Markov-Chain Monte Carlo methods.

本研究では、Comparable Corpora を用いて対訳辞書を自動的に拡張する。この対訳の自動獲得における基本的な仮説は、対訳関係にある２つの単語は同様の文脈に現れる、ということである。我々の提案手法は従来法に対し次のような点で優れている：第一に、重要な文脈をベイズ法に基づく新しい手法によって PMI (Point-wise Mutual Information)を推定することによって検出される。第二に、単語が現れる文脈を、単純な Bag-of-words モデルに加え、係り受け構造の情報も用いることで豊かに表現する。我々はさらに、係り受け構造から得られる異なった種類の情報を、ベイズ推定に基づいて適切に重みづけする手法を提案する。

Comparable Corpora とは、二つの異なる言語で書かれた、同様の内容を持つコーパス対である。対訳コーパスと異なり、Comparable Corpora 内の各文は必ずしも対訳関係にある必要は無いため、どの分野に対しても比較的簡単に Comparable Corpora が作成できる。一方、一般的な語彙に対する対訳辞書は数多く存在するが、専門的な用語に対する対訳辞書は少なく、人手による開発は高いコストを必要とする。本論文では、原言語と目標言語の Comparable Corpora を用いて、与えられた原言語の単語（対象単語）に対し適切な翻訳を以下のように検索する：まず、原言語のコーパスにおいて、対象単語が出現する文脈から、対象単語との相関値が有意なピボット語をを抽出する。こ

こで、ピボット語とは、既存の対訳辞書中に存在する内容語である。 ビポット語と対象単語の相関関係が有意かどうかは、ベイズ法を用いて PMI を推定することによって決める。
同様に、目標言語のコーパスにおいて、それぞれの対訳候補語についても、ピボット語を抽出する。最後に、対象単語のピボット語の翻訳を介して、妥当な対訳候補語を同定する。その際、各候補語の、対象単語の対訳としての妥当性は、対象単語と候補語のピボット語がランダムにマッチングする確率に基づいて計算する。

本手法では、単語の文脈において 4 種類のピボット語を考慮する。
すなわち、1 文中で単に共起した単語（Bag-of-words モデル）、に加え、
原言語と目標言語での係り受け解析結果を利用して、単語の係り受け元、係り受け先、さらに係受け木における兄弟関係にある単語をピボット語として利用する。実験では、原言語/目標言語として日本語と英語を用いた。日本語と英語の係り受け解析構造を直接対応づけることは難しいが、いくつかのヒューリスティックルールを用いて、係り受け解析結果を変形することで、係り受け解析結果が対応可能になるようにした。実験では、我々の提案手法により、従来法に比べ有意に精度が高い翻訳単語対が得られることを示す。

本論文の後半では、さらに、異なる種類の文脈情報に対して適切な重みづけを学習することで翻訳対獲得の精度を向上させる手法を提案する。
この手法は、2 つの文脈ベクトルを線形変換した後でコサイン類似度を
計算するものであり、通常のコサイン類似度の一般化になっている。
この線形変換の適切さはベイズ的な確率モデルによって表現され、変換行列の各パラメータはマルコフ連鎖モンテカルロ法を用いて学習される。